A multimodal smartphone sensor system for behaviour measurement and health status inference

A Multimodal Smartphone Sensor System for Behaviour Measurement and Health Status Inference

Dr. Daniel Kelly, Dr. Joan Condell, Prof. Kevin Curran

Ulster University

Prof. Brian Caulfield

University College Dublin

Abstract

Smartphones are becoming increasingly pervasive in almost every aspect of daily life. With smartphones being equipped with multiple sensors, they provide an opportunity to automatically extract information relating to daily life. Information relating to daily life could have major benefits in the area of health informatics. Research shows that there is a need for more objective and accurate means of measuring health status. Hence, this work investigates the use of multi-modal smartphone sensors to measure human behaviour and generate behaviour profiles which can be used to make objective predictions related to health status. Three sensor modalities are used to compute behaviour profiles for three different components of human behaviour. Motion sensors are utilised to measure physical activity, location sensors are utilised to measure travel behaviour and sound sensors are used to measure voice activity related behaviour. Sensor fusion, using a genetic algorithm, is performed to find complementary and co-operative features. Using a behaviour feature composed of motion, sound and locations data, results show that a Support Vector Machine (SVM) can predict 10 different health metrics with an error that does not exceed a clinical error benchmark.

Keywords: Health Status, Machine Learning, Motion, Location, Sound

1. Introduction

It is widely agreed that chronic diseases are the predominant challenge to global health [1]. Chronic diseases have hugely negative effects in terms of the human suffering they cause and the burden they inflict on the socioeconomic fabric of countries [2]. The World Health Organization (WHO) stated that chronic diseases, such as cardiovascular diseases, diabetes, cancers and chronic respiratory diseases, accounted for 52% of all deaths under the age of 70 worldwide. Health outcomes are used in clinical treatment, clinical trials and other

clinical research to assess the efficacy of a chronic disease treatment. Examples of different health outcomes include mortality, hospital readmission rate and disease specific physiological measurements [3]. However, there has recently been a growing focus on health outcomes that are measured by patients, termed "patient-reported health outcomes". These measures take into account the experiences of individual patients and their response to specific treatments, aspects not considered by traditional health outcomes. Health status is a particular type of patient-reported health outcome which quantifies the impact of disease on a patients daily life [4].

Studies on the reliability of health status measurement tools indicate that, while most measures are reliable for group comparisons, measures cannot be used to assess patients on an individual basis [5]. Due to the limitations of current health status measurement tools (i.e. questionnaires), there is a need for new measurement tools which can produce more accurate and reliable measurements such that clinicians can assess health status on an individual basis. The overall aim of this work is to develop a novel objective health status measurement tool using sensor technology in the community. Modern smartphones, equipped with multiple sensors built within the common and non-invasive form factor of a mobile phone, have the potential to trace human activities at scales that were previously unattainable [6]. The aim of this work is to develop an unobtrusive smartphone sensing system which can objectively measure a persons' longitudinal behaviour and make accurate predictions about their health status based on this behaviour.

In order to accurately model the mapping between mobile sensor data and health status, a participant set, with a broad spectrum of health measurements, is required. In this work, sensor data and health status information from adults in the general population are obtained using a crowd-sourced data collection methodology via a smartphone App. A multi-modal sensor system is proposed to generate behaviour profiles describing a persons' behaviour. Motion, location and sound sensors are utilised to measure physical activity, travel behaviour and voice activity related behaviour respectively. Liang et al. indicate that measure such as physical activity, measured using smartphones, along with vocal data measured by microphones and location data could be vital for screening and predicting mental health problems [7]. Our hypothesis is that a persons' behaviour will be indicative of their overall health, and that this behaviour can be captured by analysing physical activity, vocal activity and location sensor data. To test this hypothesis, we conduct a study in which participants record their motion, sound and location patterns using a smartphone and record their health status using a self-reported questionnaire. Experiments were performed on the data to discover if, and to what extent, behaviour based features extracted from motion, sound and location sensors, could be used to predict health status.

1.1. Related Work

Kelly et al. [8] carried out a study investigating links between smartphone motion sensor data and self-rated health status. A crowd-sourced dataset was recorded which comprised accelerometer and gyroscope data for 171 participants

with an average of 114 hours of data per participant. Results showed that a Support Vector Machine (SVM) regression model could predict the 10 SF-36 self-ratings with a mean absolute error of 11.7%. This paper aims to extend the work of Kelly et al. [8] by investigating additional sensor modalities as an alternative to, or to compliment, motion sensor data. We postulate that the addition of location and sound sensors will provide further insight into components of human behaviour. Ben-Zeev et al. investigated whether smartphone data, from motion, location and sound sensors, can be used as behavioural markers for mental health measurements [9]. A total of 47 young adults recorded smartphone data for a 10 week period and completed daily ratings of stress as well as pre/post measures of depression, stress and loneliness. Results, obtained using mixed effect linear modeling, showed that geo-spatial activity, sleep and variability in geospatial activity were associated with daily stress levels. Additionally, results obtained from a penalised functional regression model showed associations between changes in depression and speech duration, geo-spatial activity and sleep duration. Changes in loneliness were associated with physical activity.

To the authors knowledge, there are no other related works specifically investigating methods to automatically predict patient reported health outcomes, such as health status, using unobtrusive smartphone sensing. However, research into Body Sensor Networks (BSNs), where sensors are placed on various parts of the human body, is quite mature and BSN has been utilised for many different health based monitoring studies [35][10]. In a review of the literature, motion sensors were one of the most common sensors used in BSN studies. One of the most common approaches to utilising motion sensors for health based studies is to perform activity recognition, and use recognised activities as a method of tracking sedentary behaviour and/or activities of daily living [11]. Traditionally, health related motion sensor based studies have been conducted in controlled conditions with patients or participants wearing specialised sensors [12][13][14][15]. While there is a large number of these health-related studies using this traditional motion sensor setup, the focus of this work is to examine unobtrusive sensing using smartphones. An expansive review of health related motion sensor based studies is therefore out of the scope of this paper.

Smartphones, now equipped with high compute power and multiple embedded sensors, are being utilized by researchers for health based monitoring [16]. Kwapisz et al. [17] developed an activity recognition system using a phone-based accelerometer to record data from 29 participants performing 6 different activities. A multilayer perceptron classifier was shown to correctly classify 92% of activities. Similarly, Wannanburg et al. [18] perform a set of activity classification experiments using 10 participants performing 5 different activities. A k-Star based classification model was shown to recognise activities with 99% accuracy. Aside from general activity recognition applications, research has also shown that smartphone based motion sensors can be utilised to infer condition specific health related information. For example, Juen et al. describe a smartphone based walking monitor for patients with Chronic Obstructive Pulmonary Disease (COPD) [19] while Kelly et al. [20] conducted a case series, performing a

preliminary investigation on differences in movement patterns of COPD patients reporting problems versus COPD patients not reporting problems. Cvetkovic et al. proposed a multi-modal system using a number of sensors such as motion, heart rate, skin temperature and galvanic skin response (GSR) recorded from a smartphone and wristband to predict activity and energy expenditure [21]. Experiments showed that the system performed with 87% accuracy when predicting activities and with a mean absolute error of 0.6 when predicting energy expenditure. Gay et al. [22] developed a multi-modal smartphone based system, using multiple external sensors such as ECG and Oximeter, to monitor the wellbeing of high risk cardiac patients. ECG data is processed by the smartphone to determine if the patient is in need of help.

One of the advantages of smartphone based sensing is that multiple sensors are available within a readily available and non-intrusive form factor. Sensors such as location, microphone, light, temperature, proximity and barometer are just some of the common sensors found in modern smartphones. Researchers have investigated how some of these different sensor modalities can be utilised in the area of health monitoring. For example, Nakano et al. [23] developed a microphone based smartphone application to record ambient sound in order to detect snoring. The overall aim of the study was to monitor patients and evaluate the severity of obstructive sleep apnea. Aside from health based monitoring, there does exist a number of works relating to extracting general behaviour related information from smartphone sensors. For example, Farrahi et al. [24] investigated the use of smartphone based location data to discover human routines that characterise an individual and groups of individuals. A total of 97 participants recorded location data, using a mobile phone, for over a 16 month period. Routines such as "going to work late" and "going home early" were automatically extract from the location patterns. In another work, Farrahi et al. [25] conducted a multimodal based experiment, integrating human proximity data, via bluetooth, with location data to mine meaningful details about human activities. Using the same dataset of 97 participants, activities such as "working from 11am-5pm with 3-5 other people" and "going out from 7-midnight alone" were automatically extracted from the data. Ling et al. [26] proposed a smartphone based behaviour sensing framework combining location and motion to classify specific behaviour contexts. Accelerometer and gyroscope data were utilised to classify six locomotion activities, using a Least-Squares SVM, with 92.9% accuracy. Outdoor and indoor location sensing were also performed, using GPS and Wi-Fi positioning respectively, and sequences of location information were combined with activity classes to classify 6 different behaviour contexts, such as "fetching coffee" and "taking a break", with an accuracy of 90.3%.

While there is a large body of research work in the general area of sensors and health/human behaviour, there exists few works dealing specifically with measurement of human behaviour for the prediction of health status. In this work, we perform an investigation into the use of smartphones sensor as a method of automatically generating behaviour observations for the purpose of health status prediction. We aim to investigate the use of location, motion and sound sensors to generate behaviour measurements and identify indicators of a

persons' health status.

2. Methods

An Android smartphone App was developed to record longitudinal motion, location and microphone sensor data. Modality specific signal processing and feature extraction techniques are applied to raw sensor data to generate behaviour profiles. Behaviour profiles are then utilised as features to train and test regression models in order to predict health status. In this Section, techniques used to process raw sensor data and generate behaviour profiles are described.

2.1. Multimodal Data Processing

A key aim of the smartphone sensing app is to automatically upload sensor information to a central server where processing, fusion and analysis can be performed on the data. The first stage of the data processing is to convert raw sensor data into hourly feature summary snapshots. The reason for this processing stage is two-fold: Firstly, due to the heterogeneous nature of the sensors, data is recorded in different frequencies. Secondly, due to the quantity of data being recorded by the different sensors, it is not feasible to upload all data for each participant. Hourly summaries are therefore computed for each sensor modality in order to produce a uniform frequency for feature summaries and reduce the overall quantity of data.

Data is initially processed on the smartphone to compute hourly summary measures describing modality specific hourly behaviour. At the end of each day, hourly summary measures are uploaded to a central server. In this section the methods used to process and extract hourly summary measure from motion, location and sound sensor data are described.

2.1.1. Location - Feature Summaries

In an article published in Science Magazine, Song et al. [27] discuss the predictability of human location mobility and conclude that predictive models, driven by predictability of human mobility, are a scientifically grounded possibility with potential impact on health and well-being. Previous work by Kelly et al. [28] showed that the routined nature of human location behaviour could be leveraged to build predictive models to infer social and demographic information about a person using Global Positioning System (GPS) and cell tower data.

In this work, we postulate that location data could similarly be leveraged to infer health related information about a person. Location predictability, or location entropy, forms the basis of location behaviour measurements for this work. In order to calculate the entropy of an individuals' location behaviour, the probability distribution of a persons' location behaviour must first be modelled. Furthermore, since the nature of location behaviour is spatiotemporal, the temporal aspect of the individuals movement must also be modelled. In order to account for both these aspects of location behaviour, a two stage location behaviour model proposed by Kelly et al. [28] is used. In the first stage, the

probability distribution of a number of different geographical areas are modeled. In the second stage, the temporal transitions of the individual between these geographical areas are computed.

Geographical areas are first identified using a hierarchical clustering technique. All location co-ordinates for a person are used to automatically identify a set of locations of interest denoted as $C = \{c_1, ..., c_K\}$ where K is the total number of locations of interest computed by the clustering algorithm. For each identified area of interest, a probability distribution is computed such that the probability of an individual being at a particular geographical area c_i, given a location point p_t, is modelled using Bayes rule defined in Equation 1. The prior probability of an individual being at a location within geographical area c_i is equivalent to the significance of the cluster: $P(c_i) = \frac{N^i}{N}$, where N^i is the number of location points in geographical area c_i and N is the total number of location points. The posterior probability $P(p_t|c_i)$ is calculated using a multivariate gaussian probability density function with dimension $k = 2$, as defined in Equation 2, where $\sum_i$ is the covariance matrix calculated from all points $p \in c_i$. Finally, the combined probability of location point p_t is defined in Equation 3 where K is the total number of clusters.

$$P(c_i|p_t) = \frac{P(p_t|c_i)P(c_i)}{P(p_t)} \quad (1)$$

$$P(p_t|c_i) = \frac{1}{2\pi^{\frac{k}{2}}\sqrt{|\sum_i|}} e^{0.5[p_t-\mu(c_i)]^T[p_t-\mu(c_i)]\sum_i^{-1}} \quad (2)$$

$$P(p_t) = \sum_{i=0}^{K} P(p_t|c_i)P(c_i) \quad (3)$$

Thus far we have discussed modeling the probability of a single location point, however, human behaviour becomes much more meaningful when observed in a temporal context. A Hidden Markov Model (HMM) is therefore used to model the temporal movement of a person in relation to detected geographical areas. HMMs are characterised by a set of N states, a state transition probability matrix A, an observation symbol probability B and the initial state distribution Π. Each of the N states are used to represent each geographic area. The observation probability for each state is modeled using the location probability model $P(c_i|p_t)$. The initial state distribution is computed from the prior probability of each geographical area $P(c_i)$. The transition probability matrix is calculated by investigating all location points p_t and the state transitions that occur from time t to time $t2$, where $t2$ is the next time that a location sample is available after t. At each time t, the geographic area which the location point p_t is most likely to belong to is defined as $C(p_t)$. For each $c_i \in C$, the transition matrix A is updated, at position $[C(p_t), i]$, by adding the probability of transitioning from $C(p_t)$ to c_i with the probability that p_{t2} belongs to location c_i. After all location points are processed, the transition probability matrix, A, is then normalised such that the sum of all transitions from a particular state is equal to one.

Using the HMM, the Viterbi Algorithm could be utilised to find the most likely hidden state sequence and associated probabilities for a given sequence of locations. However, a key aim of our location based measures is to extract measures of predictability. We therefore implement a variation of the Viterbi algorithm, originally proposed by Hernando et al. [29] and modified by Kelly et al. [28], to compute the Entropy of the hidden state sequence that best explains the observations. Using a set of locations points $O_d = \{p_{d1}, ..., p_{dL}\}$ for a day d, the Viterbi Entropy algorithm computes the Entropy $H(S|O_d)$, where S is the sequence of geographic locations which best describes the set of location points O_d. Additionally, the temporal entropy, $H(S_d[t]|O_d[t] = o_d[t])$, is computed at particular times of the day t, allowing us to measure how an individuals behaviour changes over the course of a day. For each hour h of day d we define the location summary feature as $L_{dh} = H(S_d[h]|O_d[h] = o_d[h])$ where $o_d[h]$ is the median recorded location of the person during hour h of day d. For each day d, the entire set of hourly location entropy measures, $L_d = \{L_{d0}, ..., L_{d24}\}$, are uploaded to the server. Figure 1 shows the average location entropy, and rate of change, for each hour of the day, computed from all participants who uploaded at least 1 day of location data. It can be seen that the entropy naturally increases as a day progresses. This is due to the fact the probabilities for a given hour are conditional on previous hours, resulting in a lower probability and therefore a higher entropy. While the hourly location entropy will always increase, the rate at which it increases is perhaps more interesting. The rate of change of entropy shows peaks in the morning and around 5pm, which could be explained by more people traveling to/from work at those times.

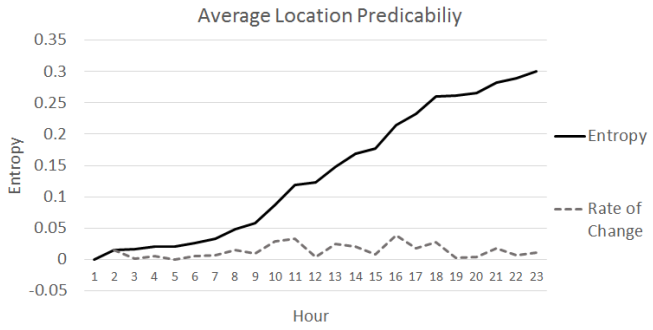

Figure 1: Average Location Predictability and Rate of Change for all participants

2.1.2. Sound - Feature Summaries

We postulate that sound activity, particularly voice activity, could be indicative of social behaviour. We therefore aim to compute sound activity profiles, for each participant, comprising of a voice activity component. The built-in microphone is utilised in order to detect sound and the presence of voice. Use of the microphone raises two possible issues however. Firstly, ethical issues around

recording a participants private conversations and secondly, power consumption issues around polling a battery depleting sensor. In order to address both of these issues, the microphone is polled every 3 minutes for a duration of 2 seconds. When polling the sensor, no actual sound information is stored and raw sound data is used only to extract sound features. Following the computation of sound features the raw sound data is immediately deleted. Features are extracted from raw sound signals using 128 ms sliding windows. For a given time t, a feature vector, a_t, is extracted in order to describe the audio characteristics of the sound at time t. Mel-frequency Cepstral Coefficients (MFCC) are used as the main set of features to characterise sound frames [30]. In addition to MFCC features, a set of time and frequency domain summary features are used, such a spectral flux, spectral centroid, bandwidth and zero crossing rate, which have been described and validated in previous works [31].

A Support Vector Machine (SVM), using a Radial Basis Function Kernel, was trained using a training set comprising 250 minutes of sounds labeled as "voice" or "other". "Voice" sounds represent recordings of different participants speaking in the foreground while "other" sounds represent background sounds from various sources such as background noises in cafes, music, traffic noise, office noise etc. Feature vectors were extracted for all labeled sounds, and an SVM was trained to classify "voice" and "other" sounds. During testing of the model, the SVM classified 95.1% of sounds correctly.

For each hour h, a total of 20 sound samples are taken. Each sample, beginning at time t, consists of 2 seconds of audio where 30 overlapping 128ms windows are utilised to extract 30 sound feature vectors $A_t = \{a_{t0}, ..., a_{t30}\}$. Additionally, sound decibel levels for each window are also recorded, denoted as $D_t = \{d_{t0}, ..., d_{t30}\}$. Features vectors which contain only silence, determined by thresholding the decibel level, are immediately classified as "other". The pretrained SVM model is then utilised to classify the remaining feature vectors, resulting in a set of sound classes $C_t = \{c_{t0}, ..., c_{t30}\}$, where $c \in \{\text{"}voice\text{"}, \text{"}other\text{"}\}$. The overall sample, A_t, is given a single classification $\bar{C}_t$ based on the percentage of individual windows classified as "voice". If more than 33% of individual classes, C_t, are classified as "voice" then the overall sample class $\bar{C}_t$ is classified as "voice". Otherwise, the overall sample class $\bar{C}_t$ is classified as "other". The mean decibel level of the sample, $\bar{D}_t$, is also recorded.

An overall sound activity summary feature S_{dh}, for hour h on day d, is computed by combining sound sample classifications and decibel level data for the 20 sound samples that were recorded during hour h. A sound activity summary feature is comprised of 3 features. The first feature is the average decibel level computed from all sample level decibels $\bar{D}_t$ for hour h. The second feature computes the fraction of samples which where not silent (i.e. above the decibel level threshold). Finally, the third feature computes the fraction of samples which where classified as "voice".

For each day d, the entire set of sound summary features, $S_d = \{\overline{S_{d0}}, \overline{S_{d1}}, ..., \overline{S_{d23}}\}$, is uploaded to the server. Figure 2 shows the average, and standard deviation, voice activity for each hour of the day, computed from all participants who uploaded at least 1 day of sound data.

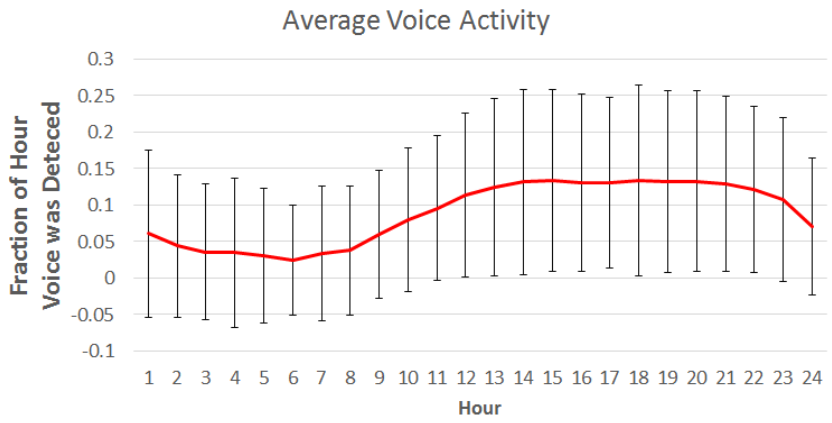

Figure 2: Average and Standard Deviation of Voice Activity for all participants

2.1.3. Motion - Feature Summaries

Physical activity, measured using motion sensors, has been shown to be a good indicator of health status. In previous work, Kelly et al. [8] propose a methodology for processing motion sensor data in longitudinal real world and uncontrolled conditions. One of the issues with utilising motion sensors for uncontrolled and longitudinal data collection is that there will often be periods of time, which we refer to as "periods of unknown", when participants are not wearing the sensor on their body. While this is not a problem for location and sound data, it is a major issue for motion sensing. It is important that these "periods of unknown" are accounted for. An "unknown" occurs when no movement is recorded from the sensor. For each unknown period, it is extremely difficult to determine if the participant is wearing the phone and being sedentary or not wearing the phone. In order to address this ambiguity, periods of unknown must be discarded. Additionally, potential movements of the smartphone in hand while the person is not moving must also be discarded. If the screen is on, we make the assumption that the phone is in hand and being used by the person. Data processing related to motion sensors must therefore only process periods of data where movement occurs and the screen is off, therefore dealing only with periods where there is a high probability that the participant is wearing the phone. It is therefore not possible to compute features based on quantifying the duration of activity, since it is possible that a participant is active during "periods of unknown". Features must therefore measure the type of movement a participant performs and not the quantity of movement performed by the participant.

Accelerometer and gyroscope sensors are used to extract 3-axis acceleration and 3-axis rotational velocity respectively. Orientation is also calculated from accelerometer and gyroscope data using the Madgwick Attitude and Heading Reference System (AHRS) [32]. Due to the unconstrained sensor placement and

orientation, a technique described by Kelly et al. [33] is utilised to transform the accelerometer and gyroscope signals into orientation independent signals by using a global reference frame to measure acceleration and rotation with respect to gravity. For each of the 8 raw motion signals, features are extracted via a series of statistical measurements performed on 2 second sliding windows. In total, a set of 82 statistical based features, denoted as M, are extracted for each 2 second window.

In order to account for "periods of unknown", only features which had a corresponding accelerometer magnitude variance greater than a pre-set threshold were used in the generation of a summary feature vector. Additionally, features which were recorded during periods when the participant was interacting with the phone were discarded. For each hour, h, all feature vectors, which have an acceleration variance greater than the threshold, were averaged to compute a single summary feature vector M_{dh} which described the overall behaviour profile of a participant for hour h on day d. For each summary feature vector M_{dh}, an associated weight, Υ_{dh} is also calculated. The weights represent the percentage of time for which the phone was moving (i.e. percentage of hour where no "periods of unknown" occurred). For each day d, the entire set of behaviour profiles, $M = \{\overline{M_{d0}}, \overline{M_{d1}}, ..., \overline{M_{d23}}\}$, was uploaded to the server along with the set of duration weights $W = \{\tilde{\Upsilon}_{d0}, \tilde{\Upsilon}_{d1}, ..., \tilde{\Upsilon}_{d23}\}$.

2.2. Behaviour Profile Generation

Thus far, we have described techniques to process raw sensor data from location, sound and motion sensors on a smartphone to generate hourly feature summaries describing a participants behaviour. Further processing of the feature summaries are performed on the server in order to generate an overall behaviour profile. The main component of the behaviour profile is computed from statistical measures of the hourly summary features. Furthermore, behaviour profiles for location and sound utilise an additional Principle Component Analysis (PCA) based feature. We denote a modality specific feature summary as $X_d = \{X_{d0}, ..., X_{d24}\}$, where $X \in \{L, S, M\}$ such that X represents one of the three feature summary vectors Location, Sound or Motion.

2.2.1. Statistical Behaviour Profile

Summary feature vectors are grouped into hourly bins, X_h^p. Where $X_h^p = \{X_{0h}^p, X_{1h}^p, ..., X_{Dh}^p\}$, and X_h^p represents all feature summaries for participant p for a specific hour h for all days from day 0 to day D and D was the total number of days recorded. A time specific average behaviour profile, which represents the average behaviour of a participant during a specific hour h for all days, was computed using a feature summary average defined in Equation 4.

$$\overline{X}_h^p = \sum_{i=0}^{D} \frac{X_{ih}^p}{D} \tag{4}$$

The average hourly behaviour profile, $\overline{X}_h^p$, represents the average behaviour of a participant at a certain time of the day. For example, $\overline{X}_{10}^p$ represents the

average behaviour of a participant between 10am and 11am for all the days a participant had the App enabled. The sequence of average hourly behaviour profiles, $\overline{X}^p = \{\overline{X}^p_0, ..., \overline{X}^p_{23}\}$, represents all hourly behaviour profiles over the course of an average day for participant p. The the overall behaviour profile $\Psi(\overline{X}^p)$, a function of the sequence of hourly behaviour profiles for participant p, is a 1-dimensional feature vector comprising a set of statistical measures applied to $\overline{X}^p$ such that $\Psi(\overline{X}^p) = \{Mean(\overline{X}^p), Var(\overline{X}^p), ROC(\overline{X}^p)\}$. Where $Mean$, Var and ROC compute the average, variance and rate of change of the sequence of average hourly behaviour profiles respectively.

2.2.2. PCA Behaviour Profile

The nature of human behaviour means that behaviour at a particular time of day is highly dependent on the behaviour that preceded it. While modeling time based behaviour is hugely important, when comparing behaviour of different individuals, measurements which are not dependent on time and behaviour which preceded it may be key in uncovering certain indicators of health. For example, two similar behaviour measurements that indicate a particular common trait, about two individuals, could potentially occur at different times of the day.

To address this limiting factor with the current representation of time, hourly feature summaries are transformed into a new space where each dimension, rather than represent a specific hour, will represent an independent variable. Each variable can then be measured independent of time and independent of other patterns in the vector. The transformation is performed on a feature specific 2-dimensional matrix $X^p(f)$, such that f represents an specific index of the feature summary vector. The feature matrix $X^p(f)$ therefore represents specific features for all hourly summary feature vectors, where columns correspond to hours and rows correspond to days. For sound, the specific feature used is $f = 2$ (fraction of hour voice was detected). For location, only one feature is recorded per hour, thus $f = 0$ (location entropy).

PCA is an orthogonal transformation technique which can transform a set of correlated variables into a set of linearly uncorrelated variables. PCA is utilised to compute a transformed feature matrix $\Omega^p_X(f)$ such that $\Omega^p_X(f) = PCA(X^p(f))$. The transformed feature matrix is then averaged by column to compute an overall PCA based behaviour profile vector $\overline{\Omega^p_X}(f)$.

2.2.3. Overall Behaviour Profile

As previously discussed, the overall motion based behavior profile is computed using only the statistical analysis of feature summaries. The overall motion behaviour profile, Γ^p_M, is therefore defined as $\Gamma^p_M = \Psi(\overline{M}^p)$. However, the overall sound and location behaviour profiles, Γ^p_S and Γ^p_L respectively, include both statistical and PCA based profiles. The overall sound behaviour profile is therefore defined as $\Gamma^p_S = [\Psi(\overline{S}^p), \Omega^p_S(2)]$, while the overall location behaviour profile is defined as $\Gamma^p_L = [\Psi(\overline{L}^p), \Omega^p_L(0)]$

2.3. Feature Fusion and Selection

Information fusion is the process of integrating data from several sources to achieve more specific inferences than could be achieved by the use of a single sensor alone [34]. Information fusion should therefore ensure that complementary and/or cooperative features are used such that additional features, from different sensors, contribute to a more complete representation of behaviour. In this work we investigate the use of data from 3 heterogeneous sensors as a means of inferring health status. While there a number of different fusion approaches, such as low-level raw data fusion and high-level decision fusion [35], we implement a mid-level feature fusion approach where the aim is to combine features from the different sensor modalities in order to produce an overall feature vector which can generate more accurate health status predictions than any of the individual sensor feature vectors. A feature subset selection process is implemented to select complementary subsets of features from the multimodal sensors. While ensuring complementary features are chosen, feature selection has additional benefits such as enhancing the generalisation of models and reducing the chances of over-fitting during training [36].

A Genetic Algorithm (GA) is utilised to perform feature subset selection [37] in order to discover subsets of complementary features from the combined set of overall behaviour profiles, $\Gamma_{LSM} = \{\Gamma_L, \Gamma_S, \Gamma_M\}$, where Γ_L, Γ_S and Γ_M represent modality specific behaviour profiles for location, sound and motion sensors respectively. A GA population is made up of a set of candidate solutions known as chromosomes. Each chromosome is a binary string, with each character corresponding to a feature. An entry of 1 signifies that the feature is selected, while 0 signifies that it is not selected. A fitness function $f(c)$ is implemented to calculate the fitness of each chromosome c in the population. The most fit chromosomes are selected for the next generations' population, and a random set of selected chromosomes are modified through a process of crossovers and mutations.

A regression model is used to train models to make health status predictions based on behaviour profiles. A fitness function $f(c, x)$ is therefore implemented to model the performance of the regression model on a specific feature subset where c is a candidate chromosome and x is the training set. During preliminary experiments, a number of different regression modeling techniques were evaluated in order to determine the best technique to carry out detailed experiments on. Results from preliminary results showed that Support Vector Machine (SVM) regression [38], using a Radial Basis Function (RBF) kernel, performed best. We therefore utilise SVMs, using a RBF, for the core experiments of this work. The fitness function $f(c, x)$ therefore computes the average Root Mean Squared Error (RMSE) for a 5 fold cross validation of training set x using a subset of features defined by c.

3. Experiments

Experiments were performed on hourly feature summaries, uploaded by participants using the App, in order to evaluate if, and to what extend, location,

sound and motion sensor data can be used to infer health status. In this Section we describe what data was collected, how it was collected and provide details of different prediction performance evaluations.

3.1. Data Collection

Participants enrolled in the study by downloading the App ("Health-U") onto their Android Smartphone via Google Play. After downloading and launching the App for the first time, participants were shown a participant consent screen where details about the study, and data collected during the study, were explained. Participants were then given the choice to consent via a button labeled "I Consent" or to reject via a button labeled "Do not participate". Ethical approval for this study was granted by Ulster University Ethics committee and the contents of the participant consent screen were reviewed by the Ethics Committee. As detailed in Section 2, a smartphone App was developed to record motion, location and microphone sensor data throughout the day and apply different signal processing and feature extraction methods to the raw data. Extracted hourly feature summaries were uploaded to a remote central storage database. To improve user retention within the experiment, functionality other than sensor recording was added to the App to provide users with visual feedback on the duration and intensity of their activities over time using graphs and statistics (see Figure 3).

Figure 3: "Health-U" App - (Left) Visual feedback showing current activity, (Middle) Activity history showing daily activity, (Right) Health Status Questionnaire.

In addition to recording sensor data, the "Health-U" App was designed to include a health status measurement tool in order to record participant health status. SF-36, a non-illness specific health status measurement tool which has been validated in a general adult population [39] and in a chronic illness patient population [40, 41], was chosen as the measurement tool for this study. The

SF-36 is a 36 question general health instrument that measures eight health related concepts: physical functioning (PF-10 items), role limitations due to physical problems (RP-4 items), bodily pain (BP-2 items), general health perceptions (GH-5 items), vitality (VT-4 items), social functioning (SF-2 items), role limitations due to emotional problems (RE-3 items), and perceived mental health (MH-5 items). Each question has multiple choice answers, with each answer having a predefined numerical score between 0-100. Answers relating to positive health contribute to a higher score, while answers relating to negative health contribute to a lower score. Each of the eight component scores are then computed using an average of specific question scores related to that component. Z-scores are then computed for each of the eight component scores and combined using weighted averages to compute two summary component measures: the Physical (PCS) and Mental (MCS) Component Summary Scores [42]. Both summary scores, PCS and MCS, are computed such that the mean and standard deviation, of a set of scores in a population, are 50 and 10 respectively.

In order to calculate the minimum sample size required for our study, a confidence interval of 95% (z-score=1.96) is used in conjunction with the largest standard deviation reported for the different SF-36 scores. Burholt et al [43] report the largest standard deviation for the RP component with a standard deviation of 32.3. Assuming a conservative margin of error of 5% in the SF-36 responses, we calculate a minimum sample size of $((1.96 \times 32.3)/2)^2 = 160$. A questionnaire screen was integrated into the App to allow participants to answer the SF-36 (See Figure 3(Right)). The App was downloaded by over 3000 users, of which 1133 completed the SF-36 questionnaire. Of these, 195, 328 and 337 users uploaded at least 24 hours of location, sound and motion feature summaries respectively. An average of 607, 559 and 685 hourly feature summaries were uploaded per user for location, sound and motion sensors respectively. Table 1 details the mean and standard deviation SF-36 self-ratings, for the 8 different concepts and the 2 summary measures, of participants as well as mean and standard deviation SF-36 scores reported by Burholt et al. for reference [43].

3.2. Qualitative Analysis

Prior to discussing experiments conducted to measure prediction performance, we first discuss some qualitative analysis performed to gain a better understanding of the data and investigate any potential relationships between features, types of behaviour and health status. Figure 4 visualises a specific feature from motion, sound and location hourly summaries for 3 different participants. Participants A, B and C reported generally high, medium and low SF-36 scores respectively. It can be seen that participant A performs more high intensity movements while participant B is generally moving for longer periods of time. Interestingly, participant C shows the lowest level of movement when duration and intensity is taken into account. For voice activity, Participant B recorded a high level of voice activity between 10am and 8pm. Participant A shows low levels of voice activity while participant C has a moderate level of voice activity later in the day between 3pm and 1am. Participant A shows very predictable location patterns, except for a period around day 50 and day 60

	Presented N=1133		Burholt et al. [43] N=13917	
	Mean	SD	Mean	SD
PCS	49.9	8.5	N/A	N/A
MCS	49.9	8.9	N/A	N/A
PF	71.9	27.4	77.8	30.0
RP	75.0	31.1	78.3	32.3
BP	69.0	25.6	70.1	28.9
GH	56.6	21.7	66.2	24.0
VT	50.3	19.8	57.3	22.3
SF	64.8	27.9	80.2	28.1
RE	64.2	32.0	87.0	26.0
MH	58.6	21.9	74.0	19.9

Table 1: Mean (Std. Dev.) SF-36 self-ratings for Participants

where predictability increases, possible due to traveling somewhere new for a number of days/weeks. Location patterns for participant B are quite unpredictable while participant C shows very predictable location behaviour. While it is not feasible to analyse data for all participants in this manner, it can be seen for this selection of participants that some potential patterns of behaviour could be linked with health status. For example, intensity of movement appears to be linked with health status, where higher intensity is shown in participants with higher health status. While no obvious/intuitive pattern can be seen linking voice activity with health status, patterns are quite distinct for each of the three participants.

3.3. Prediction Performance Test Protocol

In order to avoid feature selection bias in the health status prediction performance evaluation, we implemented a triple k-fold (k = 5) cross-validation structure modeled after that of Filzmoser et al. [44], and also utilised in a recent study by Reynolds et al. [45]. Three nested loops where implemented: 1) an outer cross-validation loop, 2) an inner cross-validation loop which contains a GA and 3) an internal GA fitness function based on cross-validation.

Figure 5 gives a visual overview of the cross validation method used. For each iteration of the outer cross validation loop, one segment is set aside as the test group. The other four segments are considered the calibration set. The calibration set is sent to the inner cross validation loop and repartitioned into 5 segments. To ensure no bias and to reduce over-fitting, each execution of the GA uses 4 of the 5 calibration segments, denoted as x, to perform GA feature selection and uses the remaining test segment y to compute the external fitness. Fitness used to determine which chromosomes are selected after each generation is known as internal fitness. Internal fitness is calculated using a 5 fold cross validation of x where 5 SVM regression models are trained on the same feature

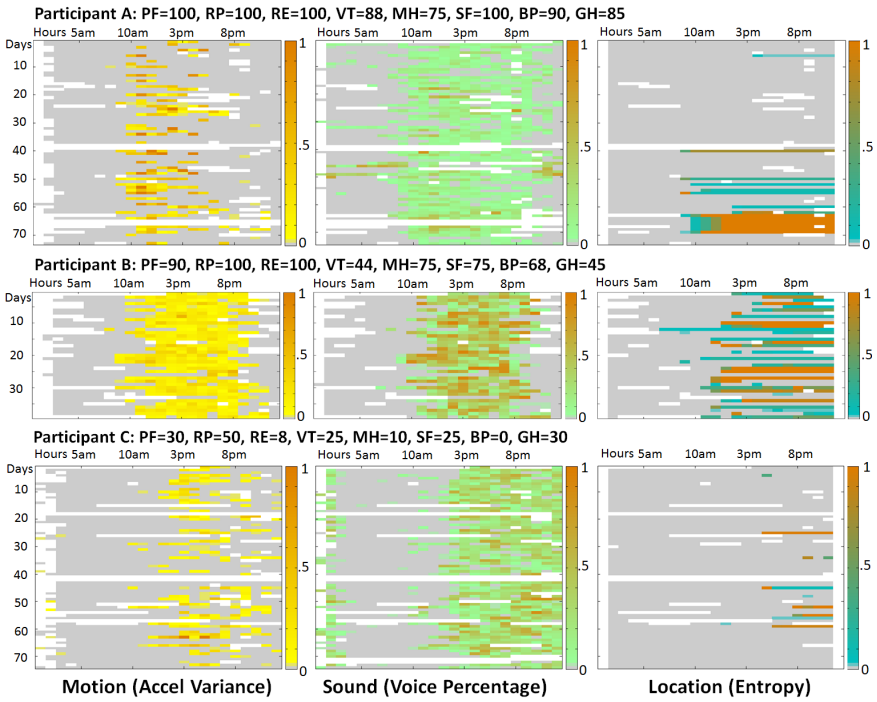

Figure 4: Visualisation of a motion, sound and location feature for 3 different participants. (White space denotes no data recorded)

subset c and 5 different subsets of x. Specifically, each SVM is trained on 4 of the 5 subsets of x and tested on the remaining subset of x. Overall internal fitness is computed from the average RMSE achieved for each of the 5 SVMs. External fitness is computed after each completed generation of the GA by training an SVM on the most fit chromosome for that generation. However, for external fitness, the model is trained on x and tested on y. The sole purpose of the external fitness is to determine at which point the GA begins to overfit to x. The GA is terminated when over-fitting is detected as determined by the generation in which the external fitness performance decreases. It is possible that a GA will overfit too early where the GA converges on a local optimum for x without ever identifying a potential global optimum. If this does occur, as indicated by no improvement in external fitness, the GA is reset and restarted.

A set of 5 individual chromosomes, C_k, are generated for each iteration of the inner cross validation loop. The individual chromosomes are then combined to create a single chromosome, $\overline{C}$, by selecting the most common features from all individual chromosomes. A feature is deemed common, if it is enabled in at least 60% of the individual feature masks. Performance for iteration i of the outer cross validation loop is calculated by training an SVM on the entire calibration set and testing on the test set, where chromosome $\overline{C}^i$ defines the subset of features to be used in training and testing. Overall performance is then calculated as the average of all performance measures.

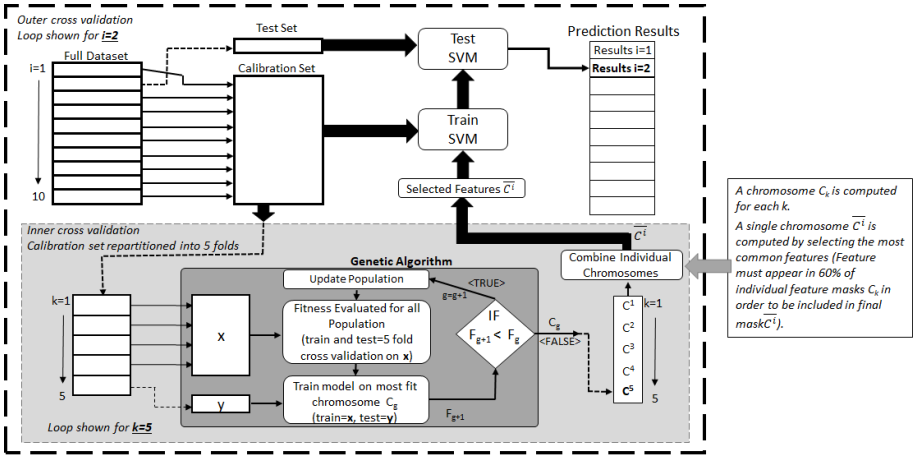

Figure 5: Feature Selection, Training and Testing Protocol Overview for Cross-Validation

3.4. Prediction Performance Results

Previous work by Kelly et al. showed that motion sensor data can be use to infer SF-36 scores with reasonable accuracy [8]. However, it is unclear whether additional sensors such as sound and location can be used to infer SF-36 measures and whether they can contribute to improved SF-36 inferences. In order

to further our understanding of how the separate sensors can indicate SF-36 measure, experiments were first performed separately on location, sound and motion only behaviour profiles, Γ_L, Γ_S and Γ_M respectively. Following this, experiments were then performed on combined location/sound, location/motion, motion/sound and location/sound/motion behaviour profiles Γ_{LS}, Γ_{LM}, Γ_{MS} and Γ_{LSM} respectively.

Evaluation of each behaviour profile is conducted by performing the triple cross validation procedure described in Section 3.3. For each behaviour profile, feature fusion and selection is performed by a GA as discussed in Section 2.3. Three evaluation metrics are calculated to evaluate performance. Pearson correlation (ρ) is used to measure the linear correlation between predicted health status and ground truth health status. Secondly, Mean Absolute Error (MAE) is used to calculate the average absolute difference between predicted health status and ground truth health status. Finally, Relative Absolute Error (RAE) computes the MAE as a percentage of the standard deviation of the health status measure. Table 2 details these performance measures for individual location, sound and motion only behaviour profiles. Of the individual behaviour profiles, it can be seen that motion performs best with an average correlation and MAE of 0.679 and 12.19 respectively. Previous work by Kelly et al. [8], report an average correlation and MAE of 0.686 and 11.72 respectively for motion sensors inferring SF-36. While additional participants are used in the current study, a paired two tailed t-test showed no statistically significant difference, for correlation ($p > 0.99$) or MAE ($p = 0.35$), between the set of 10 SF-36 measures in the current study and the previous work by Kelly et al. [8].

Initially one might postulate that sound might perform best when inferring measures related to social function (SF). However this experiment indicates otherwise with results showing that motion performs with consistently higher correlations for all measures, including SF, when compared to sound and location. We postulate that this is likely due to more noise being present in the sound related measures compared to motion. Sound measures where designed to measure duration of speech activity detected, however there exists a limitation with the approach. Speech sounds emitted by televisions, radios or computers in the vicinity of the phone could also be incorrectly detected as speech.

Table 2 details the performance measures for combined location/sound, location/motion, motion/sound and location/sound/motion behaviour profiles. Results show that combined behaviour profiles, using location, sound and motion sensors, achieves the best overall performance with correlation and MAE of 0.771 and 10.9 respectively.

Statistical significance of the prediction results, obtained from the 10 outer cross validation loops for each SF-36 measure, were evaluated in order to determine the likelihood that different performances were as a result of different sensor combinations being used. We define the null hypothesis H_0 stating that the performance measured for all sensor combinations are the same. Analysis showed that there was a statistically significant difference between sensor combinations, as determined by one-way ANOVA conducted on RAE performance ($F(6, 693) = 114.5, p < .001$) and correlation performance ($F(6, 693) =$

	Location *N=186*			Sound *N=186*			Motion *N=186*		
Measure	ρ	**MAE**	**RAE**	ρ	**MAE**	**RAE**	ρ	**MAE**	**RAE**
PCS	0.495	5.74	57.4%	0.613	5.49	54.9%	0.728	4.59	45.9%
MCS	0.447	6.48	64.8%	0.571	6.02	60.2%	0.705	5.09	50.9%
PF	0.452	17.1	62.5%	0.561	16.2	59.4%	0.724	13.4	49.2%
RP	0.351	20.1	64.6%	0.602	17.4	55.9%	0.658	15.6	50.4%
RE	0.478	20.6	64.4%	0.493	20.5	64.1%	0.743	14.4	44.9%
VT	0.475	14.1	71.6%	0.498	13.8	70%	0.633	13	65.9%
MH	0.409	16.35	74.6%	0.512	15.6	71.6%	0.611	14.1	64.7%
SF	0.421	20.6	74%	0.529	17.9	64.2%	0.681	15.84	56.7%
BP	0.348	19.2	75%	0.545	17.1	66.9%	0.644	14.6	57.3%
GH	0.439	15.5	71%	0.611	13.6	62.6%	0.737	10.9	50.3%
Average	0.432	15.6	68.1%	0.544	14.4	63%	0.686	12.19	53.7%

Table 2: Regression Prediction Evaluation Metrics for Individual Sensors

	Location/Sound *N=186*			Location/Motion *N=186*			Motion/Sound *N=186*			Location/Sound/Motion *N=186*		
Measure	ρ	**MAE**	**RAE**	ρ	**MAE**	**RAE**	ρ	**MAE**	**RAE**	ρ	**MAE**	**RAE**
PCS	0.747	4.58	45.8%	0.81	3.92	39.2%	0.760	4.36	43.6%	0.822	3.92	39.2%
MCS	0.695	5.22	52.2%	0.743	4.78	47.8%	0.720	5.3	53%	0.766	4.81	48.1%
PF	0..723	12.9	47.1%	0.771	12.3	45.1%	0.745	12.8	46.9%	0.823	11	40.2%
RP	0.701	15.2	48.8%	0.736	13.8	44.6%	0.761	13.9	44.6%	0.763	14.1	45.3%
RE	0.632	17.7	55.3%	0.772	14.2	44.6%	0.718	15.3	48.1%	0.735	15.2	47.8%
VT	0.674	12	60.6%	0.756	11.3	57.5%	0.709	12.1	61%	0.759	11.4	57.9%
MH	0.639	13.5	61.8%	0.671	13.5	61.6%	0.678	13	59.3%	0.727	12.1	55.3%
SF	0.638	16.2	58.2%	0.782	13.4	48.3%	0.723	14.1	50.4%	0.762	13.8	49.6%
BP	0.627	14.4	56.2%	0.713	13.8	54%	0.727	13.3	52.2%	0.738	12.4	48.7%
GH	0.732	11.9	55%	0.776	10.2	47.3%	0.797	9.7	44.8%	0.808	9.83	45.3%
Average	0.681	12.3	54.2%	0.754	11.1	49%	0.734	11.4	50.4%	**0.771**	**10.9**	**47.7**%

Table 3: Regression Prediction Evaluation Metrics for Sensor Combinations

C1	C2	Best (RAE)	Best (Corr)
LSM	M	LSM ($p < .001$)	LSM ($p < .001$)
LSM	S	LSM ($p < .001$)	LSM ($p < .001$)
LSM	L	LSM ($p < .001$)	LSM ($p < .001$)
LSM	ML	= ($p = .901$)	LSM ($p = .034$)
LSM	MS	= ($p = 134$)	LSM ($p < .001$)
LSM	LS	LSM ($p < .001$)	LSM ($p < .001$)
M	S	M ($p < .001$)	M ($p < .001$)
M	L	M ($p < .001$)	M ($p < .001$)
M	ML	ML ($p = .001$)	ML($p < .001$)
M	MS	=($p = .120$)	MS ($p < .001$)
M	LS	= ($p = .967$)	= ($p = .994$)
S	L	S ($p < .001$)	S ($p < .001$)
S	ML	ML ($p < .001$)	ML ($p < .001$)
S	MS	MS ($p < .001$)	MS ($p < .001$)
S	LS	LS ($p < .001$)	LS ($p < .001$)
L	ML	ML ($p < .001$)	ML ($p < .001$)
L	MS	MS ($p < .001$)	MS ($p < .001$)
L	LS	LS ($p < .001$)	LS ($p < .001$)
ML	MS	= ($p = .803$)	= ($p = .475$)
ML	LS	ML ($p < .001$)	ML ($p < .001$)
MS	LS	MS ($p = .007$)	MS ($p = .007$)

Table 4: Statistical Significance - Performance Difference between Sensor Pairs

$661.9, p < .001$). The null hypothesis was therefore rejected showing that different sensor combinations result in different SF-36 prediction performance.

A Tukey post hoc test was also performed in order to investigate performance differences between pairs of sensor combinations. Table 4 details the statistical significance of performance differences between pairs of sensor combinations as defined by the Tukey post hoc test. For RAE performance, results show that Location/Sound/Motion, Location/Motion and Sound/Motion sensor combinations result in statistically significant improvements over all other combinations, while no statistically significant difference is found among the Location/Sound/Motion, Location/Motion and Sound/Motion sensor combinations ($LSM = MS = ML$). Results of correlation performance analysis show that the Location/Sound/Motion sensor combination results in statistically significant improvements over all other combinations. Sound/Motion and Location/Motion combinations are shown to have no statistically significant difference.

Interestingly, it can be seen that while Location on its own shows no statistically significant improvements compared to any other sensor combination, Location combined with Motion results is one of the best performing sensor combinations.

3.5. Clinical Context

While results in the previous section detail performance measures and statistically significant differences between different sensor combinations, ultimately performance must consider accuracy, and error, relative to the clinical meaning of the SF-36 measures. In the literature, clinical meaning of SF-36 scores is evaluated using a benchmark. The benchmark, which is used to evaluate whether differences between SF-36 measures actually matters in terms of clinical difference, is referred to as Clinically Important Difference (CID) or Minimal Clinically Important Difference (MCID). For example, a treatment group in a clinical trial might show a 10 point SF-36 difference in physical functioning (PF) when compared to a control group, however this 10 point difference must be compared to the MCID of PF to evaluate if the difference has clinical meaning. Research has shown, based on a systematic review of 38 studies using different health status measurement tools, that the MCID was consistently close to half of a standard deviation of the health status measure [46][47]. Half a standard deviation equates to approximately 5 points for the SF-36 component scores (PCS and MCS) and approximately 10-16 points for individual SF-36 concepts (see Table 1). This has been further backed up in the literature for SF-36 self-ratings, where approximately 10, 20 and 30 points have been suggested to represent a small, moderate and large CID respectively for COPD patients for the 8 individual SF-36 self-ratings [48].

In terms of experiment results reported in the previous section, we must consider the benchmarks (CID and MCID) when evaluating the overall performance. With the literature indicating that the MCID is consistently close to 50% of a standard deviation, the RAE results are of particular interest as they report error relative to standard deviation. Results show an average RAE of 47.7% for the Location/Sound/Motion sensor combination. The implications of this is that the average SF-36 prediction error, using Location/Sound/Motion sensors, will not cause a misinterpretation in terms of clinical meaning.

A further analysis of prediction errors relative to MCID was performed, using Location/Sound/Motion sensors combination, in order to investigate the number of predictions that could be misinterpreted for clinical meaning. The total number of prediction errors which were less than the MCID were calculated for each SF-36 score. Results of this analysis showed that on average, for all 10 SF-36 scores, 63% of predictions were less that the MCID benchmark of half of a standard deviation. Furthermore, of the the remaining 37% of predictions, 17% were between half and three-quarters of the standard deviation while 10% were between three-quarters and a full standard deviation. For the sample used in this work, 37% of predictions result in a score which could be misinterpreted as a clinically relevant change. However, the scale of this misinterpretation would only be small (50%-75% of SD) for 17% of the predictions, moderate (75%-100% of SD) for 10% of prediction and large (>100% of SD) for the remaining 10%.

While additional experiments are required on a larger sample size and on different patient groups, the overall prediction performance achieved by techniques discussed in this paper is a significant result and one which indicates that SF-36

Measure	$\leq$ 50% SD	50% - 75% SD	75% - 100% SD
PCS	73.5%	16%	5.5%
MCS	58.7%	16.7%	11.8%
PF	65.7%	17.4%	10.4%
RP	68.5%	17.4%	2.1%
BP	62.2%	17.4%	10.4%
GH	64.3%	16.1%	11.8%
VT	51.7%	18.8%	11.8%
SF	64.3%	19.5%	6.9%
RE	64.3%	16.7%	11.8%
MH	54.5%	16.7%	13.2%
Average	62.7%	17.3%	9.5%

Table 5: Percentage of prediction Errors less than half of a standard deviation for Location/Sound/Motion sensor combination.

can be measured using smartphone sensors and that the predicted measures have potential to be used to make clinical interpretations without significant error.

4. Discussion

One of the aims of this work was to investigate the ability of a multi-modal sensor system to make health status predictions. Performance of the individual sensors, and combinations of sensor, reveal some interesting results. Looking at each sensor in isolation, motion related features have greater ability to predict health status compared with location and sound, while sound related features have greater prediction ability compared to location. However, motion combined with location and motion combined with sound perform with similar prediction performance. Thus, while sound performs better in isolation compared to location, sound and location provide similar levels of complementary and co-operative information to motion data. While RAE performance does not see a statistically significant improvement over any other sensor combination, combining location, sound and motion sensors does produce an overall better performing predictor compared to all other sensor combinations due to a significant improvement in prediction correlation. Overall, results comparing different sensor combinations show that a multi-modal approach to health status prediction, using location, sound and motion sensors, is a valid approach and one in which improves performance significantly when compared to a single sensor approach.

Looking at individual SF-36 score, PCS, PF and GH are measures which clearly result in the best prediction performances based on experiment results. For PCS and PF, location and motion appear to be the key sensors required for accurate prediction while motion and sound appear to be the key sensors required for accurate GH prediction. While it is perhaps clear why PCS and PF

perform well, due to these scores directly relating to physical activity and due to motion and location sensors measuring physical activity well, the accurate prediction of GH is an interesting result. GH is measured by asking participants 5 questions relating to their opinion of their own health such as "getting sick a little easier than others" and "expect my health to get worse". While there is no obvious or intuitive reason why GH is predicted with such high accuracy levels, relative to some other measures, we can only postulate that the general opinion a participant has about their own health translates into some specific type of movement and voice behaviour.

5. Conclusion

Health status has become a key measurement tool used by clinicians to assess the impact of disease on a patients daily life and to assess the efficacy of different treatments. Current measurements, using questionnaires, are limited due to their subjective nature and cannot be used to assess patients on an individual basis. New accurate and objective methods for measuring patient health status are therefore required. In this paper, we investigate the use of smartphone based multimodal sensors to measure behaviour and make objective heath status predictions based on measured behaviour. Three methods of summarising raw data from location, sound and motions sensors are proposed. Measures of location predictability are used to describe location based behaviour. Hourly statistics on voice activity and sound levels are used to describe sound based behaviour. Finally, hourly measures of motion are used to describe the type of movement performed during each hour. Hourly feature summaries are then combined into an overall behaviour profile for each participant. Sensor fusion is considered through a mid level feature fusion process using a genetic algorithm feature selection system. A crowd-sourced dataset was used to conduct experiments using a total of 186 participants. Results show that utilising all three sensor produced the overall best prediction performance for all 10 SF-36 measures with an average RAE of 47.7% and an average correlation of 0.771. Taking the MCID into account, error rates for all SF-36 measures, except VT and MH, were below the suggested benchmark of half a standard deviation. This work builds on previous work by Kelly et al. [8] and results in both statistically significant and clinically important prediction performance improvements compared to using motion sensors alone. In particular, the location and sound behaviour profiles can be used in conjunction with motion behaviour profiles, via a sensor fusion process, to produce a complimentary and co-operate set of features which can achieve more accurate inferences than can be achieved by any of the sensor alone. While additional research in terms of patient based trials is needed, the health status prediction results reported in this paper are significant and show that health status can be objectively measured using sensors. Moreover, inexpensive, unobtrusive and readily available sensors can be used to make the health status predictions.

References

[1] U. E. Bauer, P. A. Briss, R. A. Goodman, B. A. Bowman, Prevention of chronic disease in the 21st century: Elimination of the leading preventable causes of premature death and disability in the USA, The Lancet 384 (9937) (2014) 45–52.

[2] WHO, Global status report on noncommunicable diseases 2014, World Health (2014) 176.

[3] J. Curtis, D. Patrick, The assessment of health status among patients with COPD, European Respiratory Journal 21 (Supplement 41) (2003) 36S–45s.

[4] J. W. H. Kocks, M. G. Tuinenga, S. M. Uil, J. W. K. van den Berg, E. Ståhl, T. van der Molen, Health status measurement in COPD: the minimal clinically important difference of the clinical COPD questionnaire., Respiratory research 7 (i) (2006) 62.

[5] B. Gandek, J. E. Ware, N. K. Aaronson, J. Alonso, G. Apolone, J. Bjorner, J. Brazier, M. Bullinger, S. Fukuhara, S. Kaasa, A. Leplège, M. Sullivan, Tests of data quality, scaling assumptions, and reliability of the SF- 36 in eleven countries: Results from the IQOLA Project, Journal of Clinical Epidemiology 51 (11) (1998) 1149–1158.

[6] S. Majumder, M. J. Deen, S. Majumder, M. J. Deen, Smartphone Sensors for Health Monitoring and Diagnosis, Sensors 19 (9) (2019) 2164.

[7] Y. Liang, X. Zheng, D. D. Zeng, A survey on big data-driven digital phenotyping of mental health, Information Fusion 52 (2019) 290–307.

[8] D. Kelly, K. Curran, B. Caulfield, Automatic Prediction of Health Status using Smartphone Derived Behaviour Profiles, IEEE Journal of Biomedical and Health Informatics 2194 (c) (2017) 1–1.

[9] D. Ben-Zeev, E. A. Scherer, R. Wang, H. Xie, A. T. Campbell, Next-generation psychiatric assessment: Using smartphone sensors to monitor behavior and mental health., Psychiatric Rehabilitation Journal 38 (3) (2015) 218–226.

[10] G. Fortino, S. Galzarano, R. Gravina, W. Li, A framework for collaborative computing and multi-sensor data fusion in body sensor networks, Information Fusion 22 (2015) 50–70.

[11] F. Attal, S. Mohammed, M. Dedabrishvili, F. Chamroukhi, L. Oukhellou, Y. Amirat, Physical Human Activity Recognition Using Wearable Sensors, Sensors 15 (12) (2015) 31314–31338.

[12] K. Hung, Y. T. Zhang, B. Tai, Wearable medical devices for tele-home healthcare., Conference of the IEEE Engineering in Medicine and Biology Society. 7 (2004) 5384–7.

[13] D. J. Cook, M. Schmitter-Edgecombe, P. Dawadi, Analyzing Activity Behavior and Movement in a Naturalistic Environment Using Smart Home Techniques, IEEE Journal of Biomedical and Health Informatics 19 (6) (2015) 1882–1892.

[14] S. Del Din, A. Godfrey, L. Rochester, Validation of an Accelerometer to Quantify a Comprehensive Battery of Gait Characteristics in Healthy Older Adults and Parkinson's Disease: Toward Clinical and at Home Use, IEEE Journal of Biomedical and Health Informatics 20 (3) (2016) 838–847.

[15] B. R. Greene, A. Odonovan, R. Romero-Ortuno, L. Cogan, C. N. Scanaill, R. A. Kenny, Quantitative falls risk assessment using the timed up and go test, IEEE Transactions on Biomedical Engineering 57 (12) (2010) 2918–2926.

[16] S. D. Nanhore, M. M. Bartere, Mobile Phone Sensing System for Health Monitoring, International Journal of Science and Research 2 (2013) 2319–7064.

[17] J. R. Kwapisz, G. M. Weiss, S. A. Moore, Activity recognition using cell phone accelerometers, ACM SIGKDD Explorations Newsletter 12 (2) (2011) 74.

[18] J. Wannenburg, R. Malekian, Physical Activity Recognition From Smartphone Accelerometer Data for User Context Awareness Sensing, IEEE Transactions on Systems, Man, and Cybernetics: Systems (2016) 1–8.

[19] J. Juen, Q. Cheng, B. Schatz, A Natural Walking Monitor for Pulmonary Patients Using Mobile Phones, IEEE Journal of Biomedical and Health Informatics 19 (4) (2015) 1399–1405.

[20] D. Kelly, S. Donnelly, B. Caulfield, Smartphone derived movement profiles to detect changes in health status in COPD patients - A preliminary investigation., Conference of the IEEE Engineering in Medicine and Biology Society. 2015 (2015) 462–5.

[21] B. Cvetković, R. Szeklicki, V. Janko, P. Lutomski, M. Luštrek, Real-time activity monitoring with a wristband and a smartphone, Information Fusion 43 (2018) 77–93.

[22] V. Gay, P. Leijdekkers, A Health Monitoring System Using Smart Phones and Wearable Sensors, International Journal of ARM 8 (2).

[23] H. Nakano, K. Hirayama, Y. Sadamitsu, A. Toshimitsu, H. Fujita, S. Shin, T. Tanigawa, Monitoring sound to quantify snoring and sleep apnea severity using a smartphone: Proof of concept, Journal of Clinical Sleep Medicine 10 (1) (2014) 73–78.

[24] K. Farrahi, D. Gatica-Perez, Discovering routines from large-scale human locations using probabilistic topic models, ACM Transactions on Intelligent Systems and Technology 2 (1) (2011) 1–27.

[25] K. Ferrahi, D. Gatica-Perez, Probabilistic mining of socio-geographic routines from mobile phone data, IEEE Journal on Selected Topics in Signal Processing 4 (4) (2010) 746–755.

[26] L. Pei, R. Guinness, R. Chen, J. Liu, H. Kuusniemi, Y. Chen, L. Chen, J. Kaistinen, Human Behavior Cognition Using Smartphone Sensors, Sensors 13 (2) (2013) 1402–1424.

[27] C. Song, Z. Qu, N. Blumm, A.-L. Barabási, Limits of predictability in human mobility., Science (New York, N.Y.) 327 (5968) (2010) 1018–1021.

[28] D. Kelly, B. Smyth, B. Caulfield, Uncovering Measurements of Social and Demographic Behavior From Smartphone Location Data, IEEE Transactions on Human-Machine Systems 43 (2) (2013) 188–198.

[29] D. Hernando, V. Crespi, G. Cybenko, Efficient Computation of the Hidden Markov Model Entropy for a Given Observation Sequence, IEEE Transactions on Information Theory 51 (7) (2005) 2681–2685.

[30] J. D. Deng, C. Simmermacher, S. Cranefield, A Study on Feature Analysis for Musical Instrument Classification, Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on 38 (2008) 429–438.

[31] D. Kelly, B. Caulfield, Pervasive sound sensing: A weakly supervised training approach, IEEE Transactions on Cybernetics 46 (1) (2016) 123–135.

[32] S. Madgwick, An efficient orientation filter for inertial and inertial/magnetic sensor arrays, Report x-io and University of Bristol (2010) 32.

[33] D. Kelly, B. Caulfield, An investigation into non-invasive physical activity recognition using smartphones., Conference of the IEEE Engineering in Medicine and Biology Society. 2012 (2012) 3340–3343.

[34] B. Khaleghi, A. Khamis, F. O. Karray, S. N. Razavi, Multisensor data fusion: A review of the state-of-the-art, Information Fusion 14 (1) (2013) 28–44.

[35] R. Gravina, P. Alinia, H. Ghasemzadeh, G. Fortino, Multi-sensor fusion in body sensor networks: State-of-the-art and research challenges, Information Fusion 35 (2017) 68–80.

[36] M. L. Bermingham, R. Pong-Wong, A. Spiliopoulou, C. Hayward, I. Rudan, H. Campbell, A. F. Wright, J. F. Wilson, F. Agakov, P. Navarro, C. S. Haley, Application of high-dimensional feature selection: evaluation for genomic prediction in man, Scientific Reports 5 (2015) 10312.

[37] J. Yang, V. Honavar, Feature subset selection using a genetic algorithm, IEEE Intelligent Systems and their Applications 13 (2) (1998) 44–49.

[38] C.-c. Chang, C.-j. Lin, LIBSVM : A Library for Support Vector Machines, ACM Transactions on Intelligent Systems and Technology (TIST) 2 (2011) 1–39.

[39] R. Bize, J. A. Johnson, R. C. Plotnikoff, Physical activity level and health-related quality of life in the general adult population: A systematic review, Preventive Medicine 45 (6) (2007) 401–415.

[40] F. M. Boueri, B. L. Bucher-Bartelson, K. A. Glenn, B. J. Make, Quality of life measured with a generic instrument (Short Form-36) improves following pulmonary rehabilitation in patients with COPD., Chest 119 (1) (2001) 77–84.

[41] E. Ståhl, A. Lindberg, S.-A. Jansson, E. Rönmark, K. Svensson, F. Andersson, C.-G. Löfdahl, B. Lundbäck, Health-related quality of life is related to COPD disease severity., Health and quality of life outcomes 3 (2005) 56.

[42] S. S. Farivar, W. E. Cunningham, R. D. Hays, Correlated physical and mental health summary scores for the SF-36 and SF-12 Health Survey, V.1, Health and Quality of Life Outcomes 5 (1) (2007) 54.

[43] V. Burholt, P. Nash, Short Form 36 (SF-36) Health Survey Questionnaire: normative data for Wales., Journal of public health (Oxford, England) 33 (4) (2011) 587–603.

[44] P. Filzmoser, B. Liebmann, K. Varmuza, Repeated double cross validation, in: Journal of Chemometrics, Vol. 23, 2009, pp. 160–171.

[45] J. Reynolds, W. Goldsmith, J. Day, A. Abaza, A. Mahmoud, A. Afshari, J. Barkley, E. Petsonk, M. Kashon, D. Frazer, Classification of voluntary cough airflow patterns for prediction of abnormal spirometry, IEEE Journal of Biomedical and Health Informatics (2015) 1–1.

[46] G. R. Norman, J. a. Sloan, K. W. Wyrwich, The truly remarkable universality of half a standard deviation: confirmation through another look., Expert review of pharmacoeconomics & outcomes research 4 (5) (2004) 581–585.

[47] S. S. Farivar, H. Liu, R. D. Hays, Half standard deviation estimate of the minimally important difference in HRQOL scores?, Expert review of pharmacoeconomics & outcomes research 4 (5) (2004) 515–523.

[48] K. W. Wyrwich, W. M. Tierney, A. N. Babu, K. Kroenke, F. D. Wolinsky, A comparison of clinically important differences in health-related quality of life for patients with chronic lung disease, asthma, or heart disease, Health Services Research 40 (2) (2005) 577–591.